Vocabulary Learning and Instruction
Volume 11, Issue 2, December 2022
http://vli-journal.org

Developing a Discipline-Specific Corpus and High-Frequency Word List for Science and Engineering Students in Graduate School

Suwako Uehara, Hibiya Haraki and Stuart McLean
[a, b]*The University of Electro-Communications;* [c]*Momoyama Gakuin University*

Abstract

Japanese graduate school students in the field of science and engineering need to read academic research in their second language (L2), and such tasks can be challenging. Studies showed a strong (0.78) correlation between vocabulary size and reading comprehension (McLean et al., 2020), and providing high-frequency word lists could enhance comprehension. In this work-in-progress, 1.35 million tokens of professor-recommended reading materials were used to investigate a method to create a vocabulary list that would benefit science majors in graduate school; the procedures to create a corpus and a high-frequency word list efficiently; and the steps required to create a cleaner corpus. This paper outlines a systematic literature-informed method that includes input from professors in the field; the combined use of tailored script in MATLAB and AntCont (Anthony, 2022) generated corpus and high-frequency words efficiently; and repeated comparison of original PDFs and the matching text files, then adding MATLAB script to deal with specific issues created by a cleaner text. This proposed method can be applied in other contexts to enhance the generation of high-frequency word lists.

Keywords: corpus, high-frequency word list, science majors, graduate school

1 Background

This paper is a work-in-progress that reports on developing a discipline-specific corpus and a high-frequency vocabulary list for science and engineering university majors in graduate school. The students at one national science and engineering university must read scientific articles as part of their required English courses. However, an informal survey revealed that while some students are willing to read specialized scientific papers in English, others find them too difficult. To bridge this difficulty gap, a corpus from articles recommended by science faculty and a list of their high-frequency vocabulary is being developed. By using MATLAB (Version R2022a), a paid numerical analysis platform, and AntConc (Version 4.0.5), a free lexical profiling software (Anthony, 2022), a high-frequency vocabulary list from a corpus of over one million words is being developed. The first 2,000 words from the

Please cite this article as: Uehara, S., Haraki, H., & McLean, S. (2022). Developing a discipline-specific corpus and high-frequency word list for science and engineering students in graduate school. *Vocabulary Learning and Instruction, 11*(2), 57–68. https://doi.org/10.7820/vli.v11.2.uehara

New General Service Word List (NGSL: Browne et al., 2013b) were removed, leaving a discipline-specific word list. The paper considers and reports on automated steps to create a clean corpus for the efficient generation of high-frequency words.

2 Vocabulary Knowledge and Reading Comprehension

Research in reading has demonstrated that vocabulary size is an essential factor for reading success in second language (L2). Studies showed significant correlation between vocabulary size and reading comprehension tests among English as a Foreign Language (EFL) learners of various proficiency levels indicating a strong relationship between vocabulary and reading comprehension (Laufer, 1992; Qian, 2002). According to earlier research, L2 learners need a word coverage of between 95% (Laufer, 1989) and 98% (Hu & Nation, 2000) to understand written texts. In a study by Schmitt et al. (2011), they concluded that readers of academic text should aim for 98%-word coverage to improve reading comprehension. The findings also showed a relatively linear relationship between vocabulary knowledge and reading comprehension. Consequently, a list of high-frequency English words that graduate students are likely to encounter in their academic reading materials pertinent to their field of study may be useful for improving their reading comprehension of academic papers.

3 Corpus

The corpus in this study is relatively a small corpus (1.35 million tokens) of non-annotated text files. Anthony (2020), noted that while "the current trend in vocabulary use is to use ready-built software tools and pre-compiled word lists to analyze existing and new primary corpus sources" (p. 588), it is likely that more corpus linguistic researchers will make use of custom-built programs using, for example, Python and R, and to collaborate closely with software engineers and statisticians. MATLAB is another programming language, used in this study in collaboration with a postgraduate student in the field of engineering.

4 Discipline-Specific Corpus

Discipline-specific corpus in the field of science has largely been sourced to a varying degree from textbooks and lecture notes recommended by subject instructors which display representativeness (Moudraia, 2003; Ward, 2009). Corpora can then be reduced to high-frequency words by considering frequency across documents, range, and dispersion (Coxhead & Hirsch, 2007). See Appendix A for a summary of past research.

5 Decisions for Corpus Design and High-Frequency Word List

Corpus design criteria should be maximally representative of a particular language so the data can be generalized to the language variety (Butler, 2004; Sinclair, 1991). The sampling frame is often made through recommendations of lecturing staff (see Coxhead & Hirsch, 2007; Ward, 2009). Science discipline-specific papers on corpus size range from 250,000 (Ward, 2009) to 100 million tokens, where 1.76 million is considered “relatively small” (Coxhead & Hirsh, 2007). High-frequency word lists have primarily been compiled by word families (e.g., Coxhead & Hirsh, 2007), while some studies additionally share the word types. More recently, word lists have been released by flemma’s because not all word families are equally accessible or known by the learners (McLean, 2018). Nation and Hwang (1995) made a distinction between general service vocabulary and special purpose vocabulary. Inspired by this, Coxhead and Hirsh (2007)’s pilot study extracted science-specific vocabulary outside the General Service List (GSL: West, 1953) and the Academic Word List (AWL: Coxhead, 1998). A useful high-frequency word list would therefore be one in which lists could retain or remove certain existing lists depending on the study and its research questions, or based on users’ English level.

Nation and Webb (2011) provided guidelines on generating a corpus such as proofreading text fields for errors and removing text that need to be excluded from an analysis. Very few studies mention details related to editing choices of PDF or text files when proofreading the data, and how to make the process efficient. Once a corpus is made available, it would be useful to generate high-frequency word lists using various options to accommodate for range, frequency, dispersion, and retaining or removing particular word lists, and also to overcome editing decisions when deciding which part of the text will be retained.

In this study, we received recommendations from science professors for reading material suggested for their graduate students, and we are generating a small corpus specific to the field of selected professors in Engineering Science at one science and engineering university in the Kanto region. We will generate a high-frequency flemma list by removing NGSL flemmas. Following recommendations from Baker (2006), the corpus is created by us to have a better understanding of the data set.

In this work-in-progress, we therefore attempt to investigate and outline the process to create a corpus for science and engineering students in graduate school using AntConc (Anthony, 2022) and a programming software based on guidance from literature in the field of corpus, and to achieve the processing of multiple PDF files efficiently. This paper will outline the method used, and the considerations suggested to create a suitable high-frequency word list for the possible users of the list. To achieve this aim, we attempted to process multiple PDF files using the method that follows, describe the issues that were encountered, and provide suggestions on the decisions required to create a clean corpus.

6 Purposes

The purpose of this work-in-progress is to:

1. Outline a method that can be used to create a vocabulary list that might benefit science majors in graduate school;
2. Describe considerations to create a corpus and high-frequency word lists for science and engineering students efficiently;
3. Provide suggestions on decisions that are required to create a clean corpus.

7 Method

In order to create a discipline-specific high-frequency word list, the researchers followed six major steps (See Figure 1). Furthermore, in order to automate repetitive and foreseeable time-consuming steps, a computer programer was consulted periodically.

7.1 Data Collection

Fifty-eight (58) professors in an Engineering Science department were asked to recommend academic reading materials in English and 22 provided suggestions. A total of 1,182 recommendations were received. For this work-in-progress, old papers where text extraction was not feasible and papers that could not be accessed using the institutional account (98 documents) were excluded and the remaining 330 recommendations from 10 professors from the Chemistry and Biotechnology programs were processed.

7.2 Data Sorting

The data received in various forms (PDF, Weblink, reference list) were sorted into a list and each item was given an ID number. The recommended reading materials were journals (238), magazine articles (89), two doctoral dissertations, and one book chapter. See https://tinyurl.com/corpus330 for the full list and Appendix B for a breakdown of the data set.

7.3 Pre-Processing

Catalogued PDF files were edited using PDFelements to retain the main body of text in academic papers. Author names, figures, tables, acknowledgements, and reference lists were for example cut from the PDF. MATLAB (version R2022a; https://uk.mathworks.com/products/matlab.html), a programming

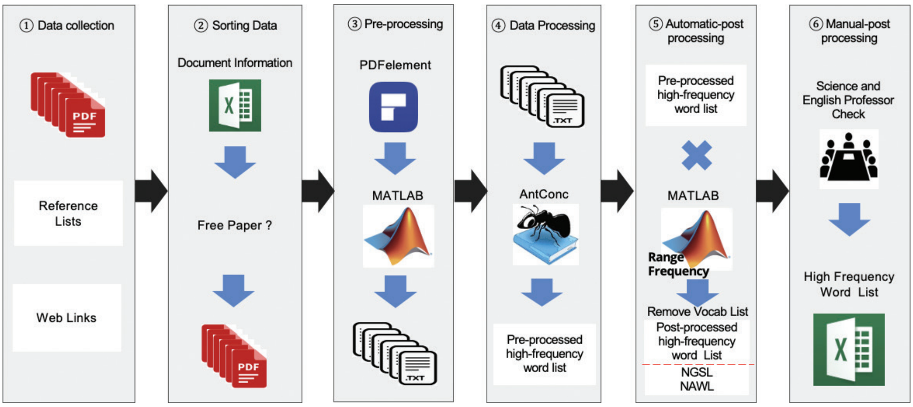
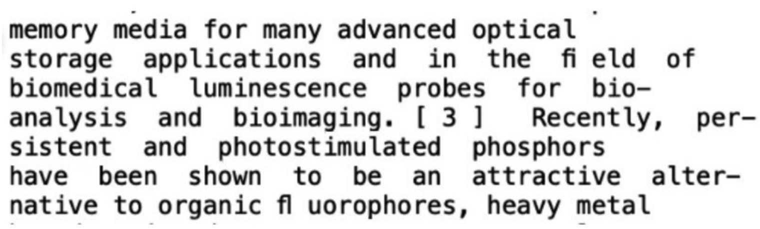


Figure 1. Process to Create a Discipline-Specific High-Frequency Word List.



```
memory media for many advanced optical
storage applications and in the fi eld of
biomedical luminescence probes for bio-
analysis and bioimaging. [ 3 ]  Recently, per-
sistent and photostimulated phosphors
have been shown to be an attractive alter-
native to organic fl uorophores, heavy metal
```

Figure 2. Ligatures and Sentence End Hyphens.

software designed to analyze and design systems using a matrix-based language that allows natural expression of computational mathematics was used to: (1) create text field from PDF, (2) fix errors that would affect word counts, and (3) add useful functions such as summary reports of word counts and processing errors, and searching for particular words in the text files, using the add-on "Text Analytics Toolbox." Typical errors when converting PDF to text files were identified through repeated and random inspection of the original PDF and the matching text files.

Two major errors were accounted for using MATLAB script (See Figure 2). A ligature is two or more characters that combine to form a single glyph that occurs with certain fonts. Scientific papers often written across numerous columns and hyphens are used to breaking up a word at the end of a line. Both of these cases pose issues when software counts words.

7.4 Data Processing

The 330 text files with an average of 4,119 words[1] per text were then processed using AntConc to obtain a high-frequency word list from 1 to 50,244 tokens, and AntConc reported a corpus size of 1,359,156 tokens.

7.5 Automatic-Post Processing

The flemma and not the lemma was used in the present study for four reasons. First, McLean (2018) found that when learners of a similar ability as the target learners demonstrated knowledge of base word forms they also demonstrated knowledge of inflectional forms. The difference between a flemma and lemma-based list is that a flemma-based list assumes that learners who can comprehend a base word form or inflectional form can also comprehend other inflectional forms, and that learners who can comprehend a word form in one part of speech (POS) can also comprehend the same word form in a different POS. Second, the nature and purpose of the list. The list being created is a Science and Engineering list and therefore the words within it are not usually base word forms or word forms with multiple POS. Third, was for practicality. When using corpus software, even if a lemma-based list is used, unless the software tags word forms for their POS then the resulting frequency list is a flemma-based frequency list. This is because unless the corpus is tagged then identical word forms of different POS are grouped and their instances are counted together. Finally, the present word list was developed to follow on from the NGSL which is a flemma-based list. The NGSL is freely available and commonly used in Japan and therefore was considered appropriate.

MATLAB scripts were made for options to remove lists per 1,000 flemmas of the NGSL, the supplementary words, and the New Academic Word List (NAWL: Browne et al., 2013a) to create a list using frequency and range chosen by the user. Also, to create a cleaner list, MATLAB was programmed to remove any one- to two-letter stings. 1,364 tokens remained with the range set to >3, frequency set to >50, and the first 2,000 NGSL flemmas set were removed.

7.6 Manual-Post Processing Science Professor Check

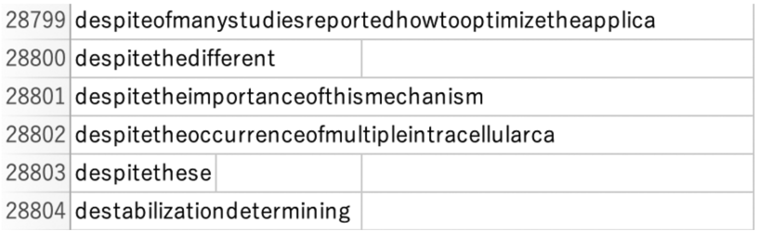
The resulting list was viewed by the professor recommending the largest number of reading materials for editing suggestions. Country and location names were removed. Singular and plural forms; present and past tense; and first and third person of the same based word was counted as the same word. The 1,364 tokens were reduced to 1,220 tokens. In the future, high-frequency words should be created with an equal volume of tokens across multiple laboratories so students can view high-frequency words across a wide range of highly specialized content. Also, some words should be left as n-grams rather than individual words (e.g., In vivo; plasma membrane).

8 Results and Discussion

The data was processed for 330 files with range set to >3 and frequency set to >50 and was set to remove the first 2,000 NGSL flemmas. It generated a high-frequency word list with 1,220 tokens. The final list from this study for Chemistry and Biotechnology majors is available here https://tinyurl.com/corpus330chembiotech. Also, see Appendix C for the sample word list.

The method to process multiple PDF files is described in the Method section earlier. The high-frequency word list that resulted from the current data set was generated heavily from one professor who recommended 188 documents. The professor's research is related to elucidating intracellular signal transduction mechanisms at the molecular level through experiments using mainly living cells of mammalian oocytes. A biochemistry department professor suggested that lists related to one specific research lab would be most beneficial for the students in their laboratory.

When building a corpus, proofreading and generating a clean corpus is one of the most time-consuming processes. To reduce the manual workload, editing of the text was autonomized using MATLAB script. MATLAB was used to edit problematic issues related to ligatures and hyphens. MATLAB was also used to remove two letter words, and select NGSL lists to be removed. The program is also equipped with the NAWL, therefore high-frequency word lists with selected NGSL and NAWL vocabulary removed can be generated with ease. Tailored high-frequency lists can be generated by adding existing word lists in MATLAB. In future, other issues that occur when converting from PDF to text files can be compiled and resolutions cane be investigated. In this way, MATLAB enabled



28799	despiteofmanystudiesreportedhowtooptimizetheapplica
28800	despitethedifferent
28801	despitetheimportanceofthismechanism
28802	despitetheoccurrenceofmultipleintracellularca
28803	despitethese
28804	destabilizationdetermining

Figure 3. Words in Sequence with No Spaces.

automatic pre-processing on a need's basis. The MATLAB scripts are available at https://tinyurl.com/CALLandResearch.

To create a clean corpus, the authors randomly compared PDFs and used various combinations of programming language and tools to efficiently output text files. From the PDF of recommended papers, PDFelements and MATLAB were used in combination, and together, we successfully processed 330 PDF documents to create text files of approximately 1.35 million tokens in 90 seconds using a computer with a standard specification. MATLAB was tailored to process issues described on ligatures, and sentence-final hyphens. British and American spelling also pose difficulties. Also, it was found that some text files had words listed in sequence with no spaces in between (See Figure 3). Using MATLAB, scripts to search for texts with long strings of characters were added. Six files were found to have between 25.9% to 65.2% of the entire text with long string characters (over 15 characters) and 24 text files included long string characters with over 100 characters. Relevant erroneous files could therefore be searched with ease and edited accordingly.

The corpus generated in this study is by no means perfect. The corpus count output by MATLAB (1,346,758 tokens) and AntConc (1,359,156 tokens) are 12,398 tokens different which is approximately 1% difference. Further investigation of the word count and closer inspection of the dataset is necessary.

9 Conclusion

This paper contributes to the existing L2 literature on corpus linguistics and high-frequency word lists for graduate students in the field of science and engineering. We successfully compiled a list of high-frequency words from 1.35 million words, for a very specific discipline using MATLAB in combination with AntConc to automatize data processing. MATLAB was used to deal systematically with some issues which became evident from inspection of the corpus multiple times. Pending issues remain which require future investigation.

One of the aims of this paper was to explore an efficient method to generate high-frequency words from recommended papers. Therefore, although the breakdown of recommendations is not equally distributed, in the future, the researchers will attempt to compile a balanced data set. Using the larger data set, high-frequency word lists can be created for relative programs in a department.

MATLAB made corpus and high-frequency word list generation efficient, however, it is meaningful to have a tool that is easily accessible for users. This current system can be applied for not only corpus in science and engineering, but to any type of selected field. Future studies will aim to adapt the MATLAB script to free programming languages such as R for increased accessibility for corpus linguists.

Acknowledgments

We express our sincere thanks to Dr. Hideki Shirakawa for his helpful discussions and comments on English vocabulary related to his research field, and to the University of Electro-Communications for the partial funding of this project.

References

Anthony, L. (2020). Resources for researching vocabulary. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 561–590). Routledge. https://doi.org/10.4324/9780429291586

Anthony, L. (2022). *AntConc (Version 4.0.5)* [Computer Software]. Waseda University. Retrieved from https://www.laurenceanthony.net/software

Baker, P. (2006). *Using corpora in discourse analysis.* Continuum.

Browne, C., Culligan, B. & Phillips, J. (2013a). *The new academic word list.* Retrieved from http://www.newgeneralservicelist.org/nawl-new-academic-word-list/

Browne, C., Culligan, B. & Phillips, J. (2013b). *The New General Service List.* Retrieved from http://www.newgeneralservicelist.org

Butler, C. (2004). Corpus studies and functional linguistic theories. *Functions of Language, 11*(2), 147–186. https://doi.org/10.1075/fol.11.2.02but

Coxhead, A. (1998). *The development and evaluation of an academic word list.* Unpublished MA thesis, Victoria University of Wellington.

Coxhead, A., & Hirsh, D. (2007). A pilot science-specific word list. *Revue Française de Linguistique Appliquée, 12*(2), 65–78. https://doi.org/10.3917/rfla.122.0065

Hu, M., & Nation, I. S. P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language, 13*(1), 403–430.

Hyland, K. & Tse, P. (2009). Academic lexis and disciplinary practice: Corpus evidence for specificity. *International Journal of English Studies, 9*(2), 111–129.

Laufer, B. (1989). What percentage of text-lexis is essential for comprehension? In C. Lauren & M. Nordman (Eds.), *Special language: From humans to thinking machines* (pp. 316–323). Multilingual Matters.

Laufer, B. (1992). How much lexis is necessary for reading comprehension? In P. J. L. Arnaud & H. Béjoint (Eds.), *Vocabulary and applied linguistics* (pp. 126–132). Macmillan.

McLean, S. (2018). Evidence for the adoption of the flemma as an appropriate word counting unit. *Applied Linguistics, 39*(3), 823–945. https://doi.org/10.1093/applin/amw050

McLean, S., Stewart, J., & Batty, A. O. (2020). Predicting L2 reading proficiency with modalities of vocabulary knowledge: A bootstrapping approach. *Language Testing, 37*(3), 389–411. https://doi.org/10.1177/0265532219898380

Moudraia, O. (2003). The student engineering corpus: Analysing word frequency. In: D. Archer, P. Rayson, A. Wilson, & T. McEnery (Eds.), *Proceedings of the corpus linguistics 2003 conference* (pp. 552–561), UCREL technical paper number 16, UCREL, Lancaster University. ISBN 1862201315.

Moudraia, O. (2004). The student engineering English corpus. *ICAME Journal, 28*, 139–143.

Mudraya, O. (2006). Engineering English: A lexical frequency instructional model. *English for Specific Purposes*, 25(2), 235–256.

Nation, P. & Hwang, K. (1995). Where would general service vocabulary stop and special purposes begin? *System, 23*(1), 35–41. https://doi.org/10.1016/0346-251X(94)00050-G

Nation, P., & Webb, S. A. (2011). *Researching and analyzing vocabulary.* Heinle Cengage Learning.

Nesi, H. (2013). 21 ESP and corpus studies. In B. Paltridge & S. Starfield (Eds.), *The handbook of English for specific purposes* (pp. 407–426). Wiley Blackwell.

Qian, D. D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning, 52*, 513–536. https://doi.org/10.1111/1467-9922.00193

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal, 95*(1), 26–43. https://doi.org/10.1111/j.1540-4781.2011.01146.x

Sinclair, J. (1991). *Corpus, concordance and collocation.* Oxford University Press.

Ward, J. (2009). A basic engineering English word list for less proficient foundation engineering undergraduates. *English for Specific Purposes, 28*(3), 170–182. https://doi.org/10.1016/j.esp.2009.04.001.

West, M. (1953). *A general service list of English words.* Longman.

Appendix A

Corpus Studies in the Field of Science and Engineering by Discipline, Summary, Corpus Size, Frequency Word List Size, and Reference

Discipline	Summary	Corpus	Frequency Word List (type)	References
Science, engineering, technology and other fields	Corpus of Professional English (http://www.perc21.org/cpe_project/index.html)	100-million-word data-base of English	1,260-word families	Nesi (2013)
Science, engineering, social sciences	Academic corpus from 30 research articles, seven textbook chapters, 20 academic book reviews in each of seven disciplines; 45 scientific letters in physics and biology theses, research articles, eight Master's thesis, six doctoral dissertations, eight final year BSc thesis across six disciplines	3-million words	1,260-word families	Hyland and Tse (2009)
Engineering	Engineering Corpus (EC) 25 textbook recommendations commonly used for 3rd-4th year undergraduate students	1/4 million words	299-word list by flemma	Ward (2009)
Science	Reading materials (textbooks, lecture notes) for 1st year students across 14 science subjects (e.g., Agricultural science, Biology Chemistry, physics, Mathematics, Computer Science etc.)	1.76-million words	315-word families	Coxhead and Hirsch (2007)
Engineering	Compulsory engineering textbooks irrelevant of field specialization	2-million words	1,260-word families or 8,850 word-types	Mundraya (2006)
Student Engineering	Student Engineering English Corpus (SEEC) from compulsory engineering textbooks irrelevant of field specialization	2-million words	1,200-word families or 9,000 word-types	Moudraia (2003, 2004)

Appendix B

Breakdown of Recommended Reading Materials by Type of Reading Material

Type of reading material	*n*	Profs	Additional information about the reading materials
Journals	238	10	Research articles published by the professors that recommended them, or papers that are relevant to the professor's lab, or papers that are cited in the field. Impact factor range was 3.23–49.96.
Book Chapter	1	1	Book chapter recommend from one research lab.
Magazine articles	89	1	Short articles with reading materials from *Nature Chemistry* on the history and discovery and current uses of elements.
Doctorate Dissertation	2	1	Dissertations highly connected to research in the professors' research lab.
Total	330	10	-

Breakdown of Recommended Reading Materials per Professor

ID	Rec	Proc	Keyword in the research field	Token count
1	4	4	Applied Physiology, Muscle plasticity, microcirculation, oxygen exchange, cell membrane function	23,131
2	89	89	Nuclear and analytical chemistry, In-beam Mössbauer Spectroscopy, Mössbauer Effect, Inorganic and Analytical Chemistry, Nuclear and Radiochemistry, Quantum Beam Science	78,382
3	67	46	Chemical biology; organic chemistry; optical physics; cancer cells; proteins; peptides; non-natural amino acids	20,1391
4	45	38	Quantum Optics; controlling and detecting the quantum nature of light. Exploit novel optical science and technology	128,290
5	188	124	Elucidation of intracellular signal transduction mechanisms at the molecular level through experiments using living cells	740,457
6	5	5	DNA Chemistry	20,155
7	5	5	Innovating and applying a bio-probe modelled on firefly bioluminescence and highly-selective hydrogenation catalyst	22,288
8	3	3	Molecular mechanisms of synaptic plasticity, which is the cellular basis for memory and learning. Development techniques to control synaptic plasticity	25,117
9	12	10	Optical and electronic properties of nanoclusters	66,997
10	10	6	Movement of bacteria	52,948
Ttl	428	330	-	1,359,156

Note. ID = Professor ID; Ttl = Total; Rec = Number of recommended papers; Proc. = Number of processed papers.

Appendix C

High-Frequency Word List (Subset for the top 100 words)

Flemma List Accounting for Singular, Plural, Present Tense and Past Tense

No.	Type	Freq	No.	Type	Freq	No.	Type	Freq
1	sperm	5,074	35	dependent	807	69	crystal	566
2	oocytes	3,613	36	detected	770	70	intensity	565
3	oscillations	2,630	37	Influx	757	71	substrate	558
4	induced	2,228	38	Pulse	755	72	mitochondria	543
5	PLC*	2,029	39	Plasma	742	73	inhibited	538
6	fertilization	2,012	40	measurements	737	74	inhibitor	537
7	activation	1,976	41	injected	733	75	transition	529
8	photon	1,905	42	wavelength	733	76	progesterone	516
9	membrane	1,663	43	Ion	713	77	GFP*	507
10	calcium	1,543	44	pathway	710	78	interference	506
11	peptide	1,484	45	spectral	699	79	cytoplasm	505
12	fluorescence	1,443	46	Buffer	693	80	CAMK II*	503
13	DNA*	1,147	47	Fusion	683	81	fluorescent	502
14	transient	1,138	48	Ratio	681	82	ATP*	498
15	acid	1,104	49	FRET*	677	83	optical	494
16	obtained	1,103	50	mammalian	675	84	incubated	492
17	activated	1,102	51	emission	663	85	dynamics	490
18	kinase	1,050	52	experimental	658	86	compounds	489
19	domain	1,024	53	Laser	650	87	mediated	484
20	receptor	1,007	54	anti***	649	88	residues	482
21	found	986	55	excitation	646	89	detection	481
22	respectively	971	56	TBA*	634	90	FURA-2*	478
23	frequency	961	57	Probe	618	91	maturation	471
24	antibody	902	58	Amino	617	92	mutant	470
25	molecules	897	59	Spectra	614	93	fertilized	461
26	injection	895	60	Trigger	612	94	aptamer	458
27	intracellular	895	61	mitochondrial	610	95	enzyme	457
28	molecular	892	62	Pump	609	96	amplitude	452
29	PLCZ*	883	63	quantum	603	97	atoms	452
30	formation	869	64	STIM*	593	98	sensitive	449
31	InsP**	852	65	CatSper*	589	99	stimulation	449
32	distribution	840	66	RNA*	585	100	regulated	449
33	extracts	828	67	Species	576			
34	embryos	814	68	Phage	572			

Note. No. = High-Frequency Ranking; Freq = Frequency.

**Abbreviations:* ATP = Adenosine triphosphate; CAMK II = Calmodulin-dependent protein kinase II (CaM kinase II or CaMKII); CatSper = Cation channels of sperm (a sperm-specific calcium channel); DNA = Deoxyribonucleic acid; FRET = Fluorescence resonance energy transfer; FURA-2 = A ratiometric and sensitive indicator dye for measuring intracellular calcium; GFP = Green fluorescent protein; GST = Glutathione S-transferase; PLC = Phosphor phospholipase C; PLCZ = Phosphor phospholipase Z; RNA = Ribonucleic acid; STIM = Stromal interaction molecules; TBA = *tert*-Butyl acrylate

***Variations:* For example, InsP$_3$ = Inositol trisphosphate; InsP$_4$ = plasmalemmal inositol 1,3,4,5-tetrakisphosphate

****Prefix:* For example, Anticipated; Antibody; Anti-rabbit IgG; Anti-poptotic; Anti-Securin; Anti-CDC2